# Cerberus Transformer: Joint Semantic, Affordance and Attribute Parsing Supplementary Material

Xiaoxue Chen[1], Tianyu Liu[2], Hao Zhao[3,4], Guyue Zhou[1], Ya-Qin Zhang[1]
[1]AIR, Tsinghua University [2]HKUST [3]Peking University [4]Intel Labs
{chenxiaoxue, zhouguyue, zhangyaqin}@air.tsinghua.edu.cn
tianyu.liu@connect.ust.hk, zhao-hao@pku.edu.cn, hao.zhao@intel.com

In this supplementary material, we provide more experiment results to show the effectiveness of Cerberus and validate our assumption that task affinity helps Cerberus learn under weak supervision. In the following sections, we first provide per-category evaluation results for semantic, affordance and attribute parsing respectively. Then we provide more visualization results for both parsing and attention. Shared attention weights reveal the potential reason behind strong weakly-supervised learning performance.

## 1. Per-Category Evaluation

Tab. 1, 2 and 3 demonstrate our per-category semantic, affordance and attribute mIoU results on NYUd2 [1] respectively. *Single* denotes the separately trained model for different tasks. And the percentage represents the amount of annotation used by weak supervision. It is manifested that our Cerberus achieves the best performance in most categories. And it also achieves superior performance than *Single* under the same amount of weak supervision.

Notably, the mIoUs of certain semantic labels are 0 when using a weakly supervised *Single* model. And we observe that these categories often appear as small regions, like paper, or have a complicated internal structure, like person. For these categories, the number of randomly sampled pixels is too small to provide enough information for semantic parsing. Hence these sub-tasks cannot be learned under weak supervision effectively. However, when using a Cerberus model, the mIoUs of these sub-tasks are greater than zero, which verifies task affinity does help weakly-supervised learning, especially for those hard sub-tasks.

## 2. More Qualitative Results

We show more qualitative results in Fig. 1. We choose various indoor scenes with different semantic, affordance, and attribute labels. As shown in Fig. 1, Cerberus achieves precise attribute, affordance and semantic parsing in all these scenes. For example, in row 6 of Fig. 1, though both sides of the room are white, Cerberus can precisely distinguish the difference between blinds and walls. Accurate parsing is also observed in the *Painted* attribute results. And in the affordance results of row 6, Cerberus can even identify regions on the bed that are too far away to sit on. Considering the diversity of scenes, we believe Cerberus is accurate enough for various applications including augmented reality and intelligent service robots .

## 3. More Attention Visualization

Fig. 2 provides more attention maps and corresponding parsing results. We train three Cerberus with one task supervised by $1\%$ annotation while the other two by full supervision, and compare them with fully-supervised Cerberus. As shown in the figure, attention maps focus on those regions that correspond to parsing results. For example, in row 8 of Fig. 2, the attention weights exactly focus on all movable objects, including objects on shelves, objects on the cabinets, and even the painting on the wall. Meanwhile, for the weakly-supervised model, we still can find attention maps showing the corresponding features, which appear very similar to the fully supervised attention maps. We believe that this is because those related sub-tasks help attention learning with little annotation. And with shared attention, we achieve strong results under weak supervision.

## 4. Computation Cost and Failure case

In Fig. 4, the computation cost is visualized as the training time distribution pie chart, which shows solving optimal weights is efficient and only takes $10\%$ of the training time.

As shown in Fig. 3 for the failure case, due to the high affinity between *walkable* and *textured*, Cerberus is affected by the shared representation and performs worse than a separately trained attribute network.

## References

[1] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd

| Method | Wall | Floor | Cabinet | Bed | Chair | Sofa | Table | Door | Window | Bookshelf | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single (1%) | 71.5 | 82.0 | 53.9 | 67.4 | 59.0 | 60.8 | 46.2 | 20.8 | 43.3 | 42.4 | 23.9 |
| Single (0.5%) | 71.6 | 81.2 | 53.3 | 63.2 | 57.4 | 55.1 | 44.3 | 15.7 | 30.5 | 39.7 | 20.2 |
| Single (0.1%) | 68.3 | 80.3 | 45.1 | 64.0 | 52.2 | 58.5 | 42.8 | 5.7 | 28.6 | 38.2 | 18.6 |
| Cerberus (1%) | 77.7 | 87.3 | 63.9 | 71.0 | 64.2 | 65.4 | 51.7 | 37.5 | 50.1 | 44.3 | 42.7 |
| Cerberus (0.5%) | 75.9 | 87.0 | 63.6 | 70.7 | 65.5 | 67.5 | 50.5 | 33.1 | 51.0 | 41.9 | 39.9 |
| Cerberus (0.1%) | 79.4 | 86.8 | 62.3 | 70.1 | 64.2 | 66.2 | 46.8 | 41.7 | 48.8 | 43.0 | 39.1 |
| Single | 80.0 | 88.0 | 61.8 | 72.7 | **66.6** | **67.8** | 53.2 | **43.0** | **52.5** | 43.8 | 48.8 |
| Uniform | **80.8** | 87.7 | 62.3 | 71.7 | 65.2 | 64.1 | 49.8 | 42.3 | 50.4 | **46.0** | 48.3 |
| Cerberus | 79.6 | **88.7** | **64.7** | **72.8** | 64.7 | 65.5 | **54.4** | 41.1 | 48.0 | 44.5 | **50.4** |

| Method | Picture | Counter | Blinds | Desk | Shelves | Curtain | Dresser | Pillow | Mirror | Floormat | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single (1%) | 51.5 | 56.0 | 55.3 | 13.4 | 1.4 | 51.5 | 34.2 | 26.8 | 0.3 | 15.6 | |
| Single (0.5%) | 47.0 | 56.7 | 52.7 | 3.6 | 3.0 | 17.6 | 6.9 | 19.5 | 1.1 | 0.4 | |
| Single (0.1%) | 47.3 | 55.6 | 48.1 | 0.0 | 0.0 | 10.8 | 1.1 | 10.1 | 0.0 | 8.3 | |
| Cerberus (1%) | 58.4 | 67.0 | 64.8 | 28.2 | 14.7 | 59.0 | 47.4 | 43.5 | 42.4 | 42.3 | |
| Cerberus (0.5%) | 56.1 | 65.7 | 63.1 | 28.0 | 10.8 | 63.3 | 47.1 | 45.4 | 46.1 | 35.3 | |
| Cerberus (0.1%) | 58.1 | 64.6 | 64.2 | 26.7 | 12.8 | 60.5 | 41.1 | 43.5 | 43.3 | 40.3 | |
| Single | 58.1 | 66.7 | 61.8 | 27.0 | **17.7** | 58.4 | 47.6 | 43.1 | **49.3** | 39.8 | |
| Uniform | 57.1 | 68.8 | **65.3** | 26.6 | 16.9 | 60.0 | 51.9 | 35.2 | 49.0 | 45.9 | |
| Cerberus | **60.5** | **70.6** | 63.8 | **28.3** | **17.7** | **64.7** | **54.8** | **44.2** | 47.4 | **46.2** | |

| Method | Clothes | Ceiling | Books | Fridge | Television | Paper | Towel | S-curtain | Box | W-board | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single (1%) | 8.7 | 0.1 | 1.0 | 0.0 | 43.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | |
| Single (0.5%) | 5.5 | 0.6 | 0.9 | 0.0 | 39.6 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | |
| Single (0.1%) | 5.6 | 0.0 | 0.0 | 0.0 | 35.4 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | |
| Cerberus (1%) | 21.5 | 56.2 | 18.7 | 21.7 | 65.7 | 5.6 | 36.1 | 16.1 | 9.4 | 38.7 | |
| Cerberus (0.5%) | 21.7 | 57.4 | 8.6 | 0.0 | 61.0 | 0.2 | 36.6 | 13.5 | 3.0 | 65.6 | |
| Cerberus (0.1%) | 16.8 | 55.8 | 21.3 | 44.2 | 62.5 | 1.5 | 5.0 | 17.0 | 3.0 | 0.2 | |
| Single | **23.0** | 56.5 | **36.8** | **66.4** | 60.3 | 30.7 | 48.2 | 33.7 | 15.9 | 74.8 | |
| Uniform | 22.9 | 59.9 | 35.2 | 58.9 | 57.8 | 32.2 | 44.8 | 38.6 | 14.5 | **80.5** | |
| Cerberus | 22.4 | **60.2** | 34.8 | 64.9 | **65.5** | **38.2** | **50.3** | **45.2** | **18.7** | 74.2 | |

| Method | Person | N-stand | Toilet | Sink | Lamp | Bathtub | Bag | O-str | O-furnitr | O-prop | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Single (1%) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 9.7 | 6.9 | 33.1 | |
| Single (0.5%) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 3.3 | 2.9 | 24.5 | |
| Single (0.1%) | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 6.7 | 1.1 | 31.7 | |
| Cerberus (1%) | 74.6 | 9.1 | 61.2 | 47.3 | 34.9 | 24.7 | 0.0 | 27.3 | 20.6 | 38.9 | |
| Cerberus (0.5%) | 76.0 | 0.3 | 63.1 | 35.3 | 0.1 | 2.3 | 0.0 | 26.9 | 20.5 | 37.6 | |
| Cerberus (0.1%) | 74.4 | 0.6 | 56.6 | 42.9 | 0.0 | 9.6 | 0.0 | 27.5 | 21.4 | 37.2 | |
| Single | **82.1** | 42.8 | 64.1 | 46.8 | 42.0 | 34.9 | 0.0 | **33.4** | **22.1** | 40.2 | |
| Uniform | 70.1 | 43.1 | 62.6 | 49.7 | 40.1 | 32.9 | 0.0 | 31.6 | 20.2 | 40.0 | |
| Cerberus | 74.7 | **43.3** | **66.4** | **53.7** | **44.3** | **37.2** | **6.0** | 33.1 | 21.8 | **40.6** | |

Table 1. Per-category semantic parsing results on NYUd2.

| **Method** | Lyable | Movable | Reachable | Sittable | Walkble | mIoU |
|---|---|---|---|---|---|---|
| Single (1%) | 40.1 | 55.1 | 87.4 | 41.1 | 81.1 | 60.9 |
| Single (0.5%) | 37.1 | 53.4 | 86.2 | 36.6 | 80.6 | 58.8 |
| Single (0.1%) | 39.8 | 45.5 | 84.9 | 39.6 | 77.1 | 57.5 |
| Cerberus (1%) | 51.3 | 57.3 | 87.9 | 41.1 | 82.9 | 64.1 |
| Cerberus (0.5%) | 49.1 | 57.0 | 87.9 | 39.5 | 84.0 | 63.5 |
| Cerberus (0.1%) | 51.3 | 57.3 | 87.8 | 41.1 | 82.9 | 64.1 |
| Single | 51.4 | 57.5 | 87.7 | 43.4 | 85.9 | 65.2 |
| Uniform | 47.2 | 55.8 | 88.1 | 43.5 | 85.1 | 63.9 |
| Cerberus | **53.1** | **58.7** | **88.9** | **44.2** | **88.3** | **66.3** |

Table 2. Per-category affordance parsing results on NYUd2.

| **Method** | Wood | Painted | Paper | Glass | Brick | Metal | Flat | Plastic | Textured | Glossy | Shiny | mIoU |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Single (1%) | 46.2 | 64.3 | 22.0 | 37.8 | 45.4 | 13.2 | 0.0 | 27.7 | 72.4 | 46.4 | 46.0 | 38.3 |
| Single (0.5%) | 49.8 | 62.5 | 10.6 | 36.5 | 44.1 | 12.0 | 0.0 | 25.5 | 71.9 | 48.4 | 46.6 | 37.1 |
| Single (0.1%) | 47.4 | 63.8 | 8.5 | 37.7 | 45.6 | 11.6 | 0.0 | 25.8 | 67.0 | 46.8 | 45.9 | 36.4 |
| Cerberus (1%) | 52.0 | 67.5 | 33.2 | **45.5** | 50.1 | 21.1 | 3.6 | 30.8 | 76.2 | 51.0 | 53.9 | 44.1 |
| Cerberus (0.5%) | 52.9 | 66.8 | 34.5 | 45.4 | 50.2 | 21.0 | 4.2 | 30.4 | 75.4 | 51.2 | **54.1** | 44.2 |
| Cerberus (0.1%) | 52.4 | 67.3 | 27.0 | 45.4 | 52.5 | 21.9 | 3.6 | 35.2 | 74.8 | 49.2 | 49.7 | 43.5 |
| Single | 52.2 | 66.8 | 30.7 | 44.6 | **52.6** | 20.8 | 2.5 | **35.3** | 75.6 | 51.5 | 54.0 | 44.2 |
| Uniform | **54.5** | 67.8 | 29.1 | 43.2 | 51.7 | 25.0 | **6.2** | 31.1 | **76.4** | 50.8 | 53.8 | 44.5 |
| Cerberus | 54.3 | **68.1** | **36.2** | 45.3 | 51.9 | **25.1** | 5.4 | 31.9 | 74.5 | **51.8** | **54.1** | **45.3** |

Table 3. Per-category attribute parsing results on NYUd2.

images. In *European conference on computer vision*, pages
746–760. Springer, 2012. 1

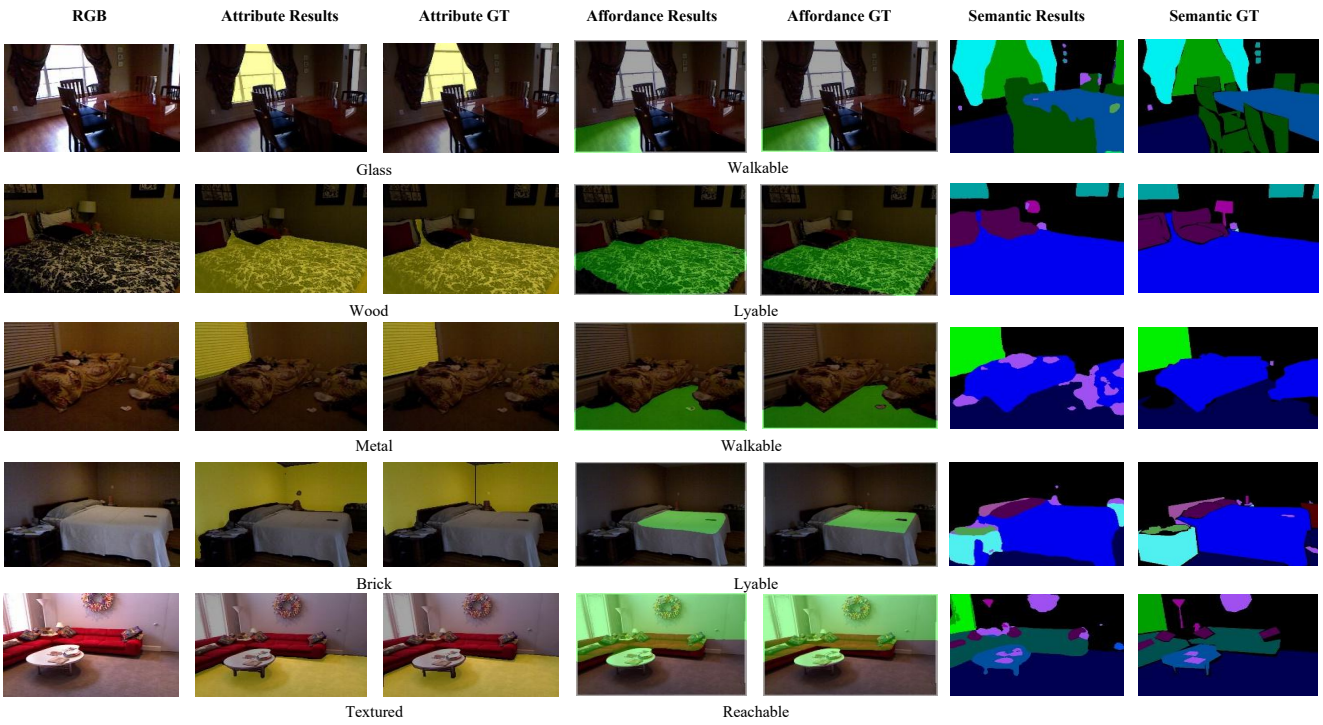


Figure 1. More qualitative results.

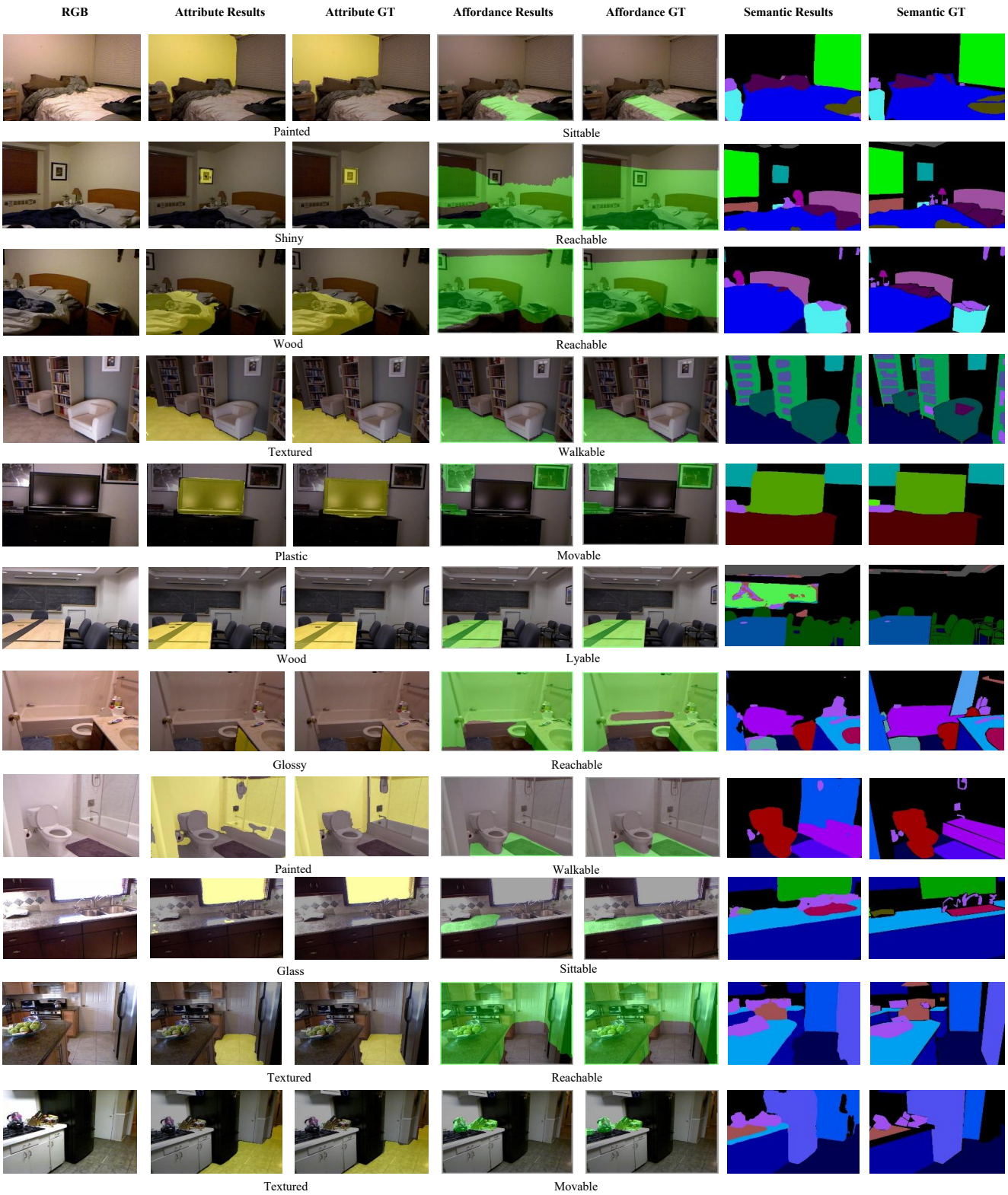


Figure 1. More qualitative results (cont.).

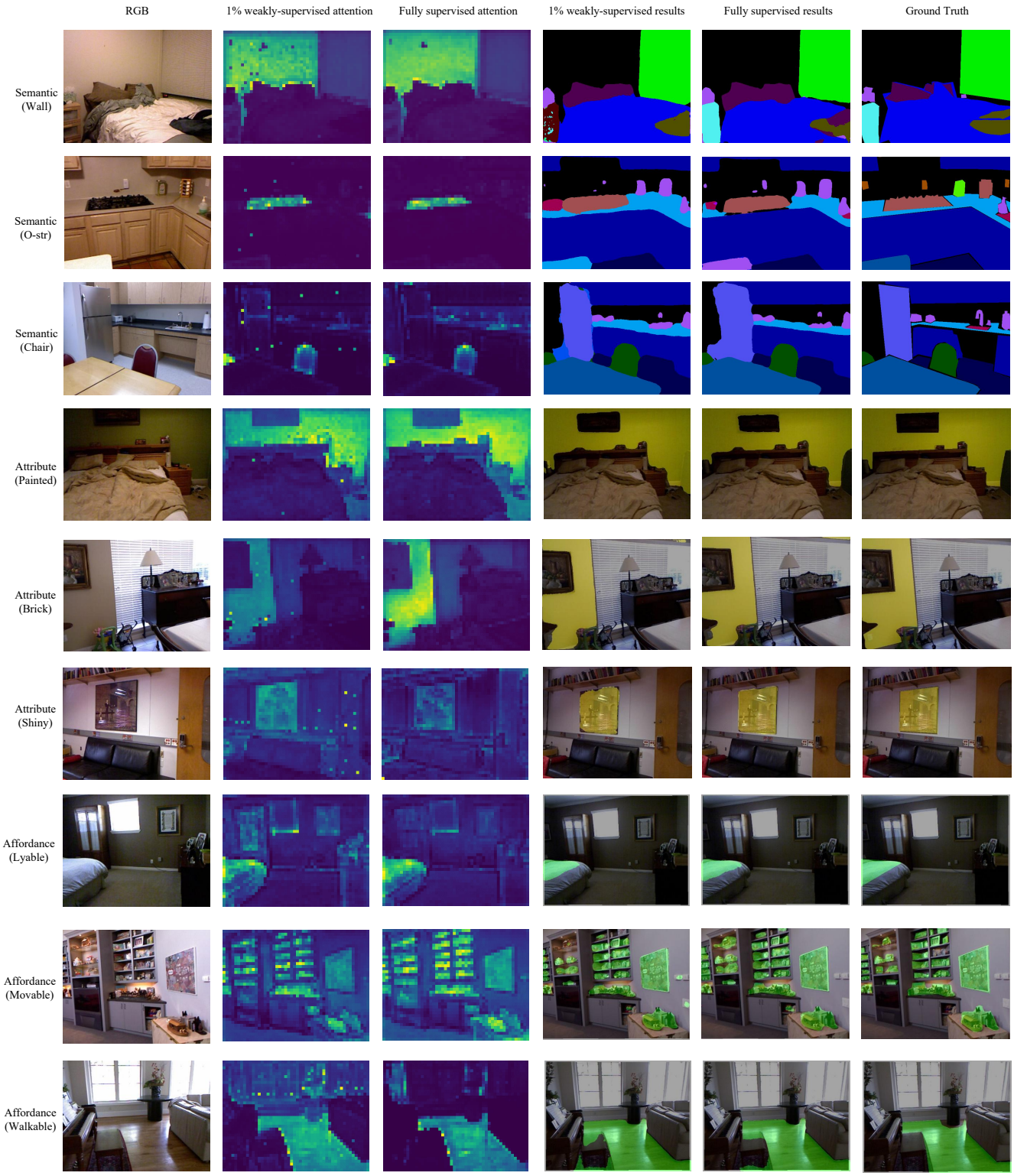


Figure 2. More attention visualizations.

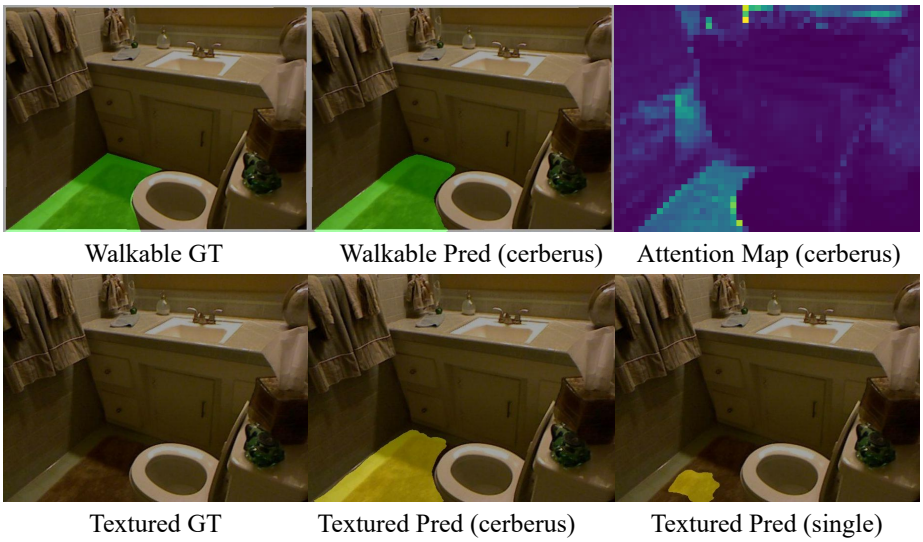


Figure 3. Failure cases.

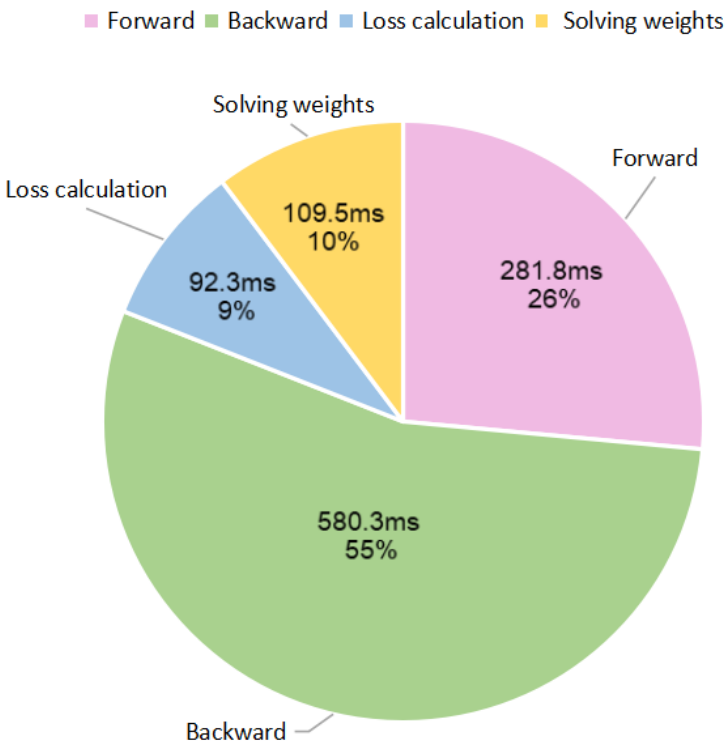


Figure 4. Computation cost distribution pie chart.